Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications xx (xxxx) xxx–xxx

stochastic processes and their applications

www.elsevier.com/locate/spa

Law of large numbers for the many-server earliest-deadline-first queue

Rami Atar[a,*], Anup Biswas[b], Haya Kaspi[c]

[a] *Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel*
[b] *Department of Mathematics, Indian Institute of Science Education and Research, Pune 411008, India*
[c] *Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 32000, Israel*

Received 11 October 2016; received in revised form 3 May 2017; accepted 12 September 2017
Available online xxxx

Abstract

A many-server queue operating under the earliest deadline first discipline, where the distributions of service time and deadline are generic, is studied at the law of large numbers scale. Fluid model equations, formulated in terms of the many-server transport equation and the recently introduced measure-valued Skorohod map, are proposed as a means of characterizing the limit. The main results are the uniqueness of solutions to these equations, and the law of large numbers scale convergence to the solutions.
© 2017 Elsevier B.V. All rights reserved.

MSC: 60F17; 60G57; 68M20

Keywords: Measure-valued processes; Measure-valued Skorohod map; Many-server transport equation; Fluid limits; Earliest-deadline-first; Least-patient-first; Many-server queues

1. Introduction

This paper proves a law of large numbers (LLN) many-server limit for a queueing model with general service time and deadline distributions, operating under the *earliest-deadline-first* (EDF) scheduling policy (we refer to this model as G/G/N+G EDF). By a *many-server limit* we refer to a setting where the number of servers grows without bound. The limit is characterized in terms of a set of so-called *fluid model equations* (FME) that involve both the *many-server transport*

* Corresponding author.
E-mail address: atar@ee.technion.ac.il (R. Atar).

https://doi.org/10.1016/j.spa.2017.09.009
0304-4149/© 2017 Elsevier B.V. All rights reserved.

UNCORRECTED PROOF

equation (MSTE) [15] and the recently introduced *measure-valued Skorohod map* (MVSM) [3]. It provides the first result on the EDF policy involving a many-server limit. Several papers have analyzed EDF asymptotically by appealing to the so-called *frontier process* (see below). However, as argued in [3], the method based on this process is not generic enough to cover a large variety of models (especially ones with time-varying parameters). Our motivation is to extend the asymptotic analysis of EDF to settings where the method involving frontier process is not expected to be effective; the many server regime offers a natural setting of this sort (even when the parameters are constant over time).

In recent years the use of measure-valued processes in mathematical modeling of queueing systems has been very successful. As far as many-server asymptotics are concerned, it is well understood since the work of Halfin and Whitt [11] that exponential service time distribution leads to simple limit dynamics; specifically, the diffusion-scale heavy traffic limit of [11] is characterized by a diffusion process on the real line. However, there is a great deal of motivation to study many-server models under general service time distributions, stemming from applications such as call centers and cloud computing. For example, the statistical study of a call center by Brown et al. [5] finds a good fit of the service time data to the lognormal distribution. In this vein, in [24], Whitt considered a G/G/N system with abandonment and proposed a deterministic LLN (otherwise referred to as fluid) approximation. In [15], Kaspi and Ramanan obtained measure-space valued fluid limits for such systems. Kang and Ramanan generalized this work by modeling customer abandonment [13]. Further generalizations in this direction were obtained by Zhang [25] and Walsh-Zuñiga [23]. In [4], Atar et al. studied multi-class many-server queues with fixed priority and established the existence of a unique fluid limit. Kang and Ramanan studied ergodic properties of the G/G/N+G model and its relation with the invariant states of the fluid limit in [14]. Reed [22] established the fluid and diffusion limits of the customer-count processes for many-server queueing systems under a finite first moment assumption on service time distribution.

The aforementioned works were concerned with the *first-come first-served* (FCFS) discipline. Many-server systems operating under the EDF discipline were considered recently by Mandelbaum and Momčilović [18], where a fluid limit heuristic was developed. Motivated by prioritizing customers with least patience and emphasizing the importance of this policy for emergency services, [18] refers to this policy as *least-patience-first*. Decreusefond and Moyal [8] study the fluid limits of M/M/1+M EDF, and Atar et al. [2] generalize these results to G/G/1+G EDF. Diffusion limits for G/G/1+G EDF systems were studied by Doytchinov et al. [9] and Kruk et al. [17]. For additional work on EDF in asymptotic regimes other than the many-server limit we refer to Kruk [16] and references therein.

An attractive feature of EDF, established in several of the aforementioned settings, is that it minimizes the abandonment count. Specifically, in [19–21] it was shown for a single server model that EDF minimizes customer abandonments within a certain class of scheduling policies. The paper [17] studies G/G/1+G and shows that the *reneged work* is minimized under EDF. Comments in the introduction of [18] also address this minimality property, and so do some of the results of Section 4.1.5 of [3].

In this article we are interested in the many-server LLN limit of the G/G/N+G EDF. To elaborate on the hurdles in obtaining fluid limits in this setting let us briefly mention some tools that have been used in the literature to treat the LLN for the single server EDF. The aforementioned frontier process has been one of the main tools in [2,9,17]. One defines the lead time of a customer at time t as the (possibly negative) difference between the customer's deadline and the time t. The frontier process at time t is defined as the maximum lead time at t

UNCORRECTED PROOF

Please cite this article in press as: R. Atar, et al., Law of large numbers for the many-server earliest-deadline-first queue, Stochastic Processes and their Applications (2017), https://doi.org/10.1016/j.spa.2017.09.009.

Download English Version:

https://daneshyari.com/en/article/7550107

Download Persian Version:

https://daneshyari.com/article/7550107

Daneshyari.com